Evidence on the measurement properties of health-related quality of life instruments is largely missing in patients with low back pain

Chiarotto, Alessandro; Terwee, Caroline B.; Kamper, Steven J.; Boers, Maarten; Ostelo, Raymond W.

published in
Journal of Clinical Epidemiology
2018

DOI (link to publisher)
10.1016/j.jclinepi.2018.05.006

document version
Publisher's PDF, also known as Version of record

Journal of Clinical Epidemiology 102 (2018) 23–37

Journal of Clinical Epidemiology

REVIEW

Evidence on the measurement properties of health-related quality of life instruments is largely missing in patients with low back pain: A systematic review

Alessandro Chiarotto[a,b,*], Caroline B. Terwee[a], Steven J. Kamper[c,d], Maarten Boers[a,e], Raymond W. Ostelo[a,b]

[a]*Department of Epidemiology and Biostatistics, Amsterdam Public Health research institute, VU University Medical Center, Amsterdam, The Netherlands*
[b]*Department of Health Sciences, Amsterdam Movement Sciences research institute, Vrije Universiteit, Amsterdam, The Netherlands*
[c]*School of Public Health, University of Sydney, Sydney, Australia*
[d]*Centre for Pain, Health and Lifestyle, Australia*
[e]*Amsterdam Rheumatology and Immunology Center, VU University Medical Center, Amsterdam, The Netherlands*

Accepted 14 May 2018; Published online 21 May 2018

Abstract

Objective: To synthesize the measurement properties of six health-related quality of life instruments (Short Form 36 [SF-36], Short Form 12 [SF-12], EuroQol 5D-3L [EQ-5D-3L], EuroQol 5D-5L [EQ-5D-5L], Nottingham Health Profile (NHP), and Patient-Reported Outcome Measurement Information System Global Health [PROMIS-GH-10]) in patients with low back pain (LBP).

Study Design and Setting: Six electronic databases (MEDLINE, EMBASE, CINAHL, PsycINFO, SportDiscus, and Google Scholar) were searched (July 2017). Studies assessing any measurement property in nonspecific LBP patients were included. Two reviewers independently screened the articles and assessed the risk of bias (COSMIN checklist). Consensus-based criteria were used to rate measurement properties results as sufficient, insufficient, or inconsistent; a modified GRADE approach was adopted for evidence synthesis.

Results: High quality evidence was found for insufficient construct validity of SF-36 summary scores, and EQ-5D-3L utility and visual analogue scale scores. Moderate evidence was found for sufficient construct validity of SF-12 physical summary score and inconsistent responsiveness of EQ-5D-3L utility score. Very low quality evidence was found on each instrument's content validity; very low to low evidence underpinned the other assessed measurement properties. EQ-5D-5L, NHP and PROMIS Global Health-10 were not evaluated in LBP patients.

Conclusion: Documentation of the measurement properties of health-related quality of life instruments in LBP is incomplete. Future clinimetric studies should prioritize content validity.

Keywords: Health-related quality of life; Measurement instruments; Measurement properties; Low back pain; COSMIN

1. Introduction

Low back pain (LBP) represents one of the most burdensome and costly health conditions [1,2]. This condition has important impacts on the patients' health-related quality of life (HRQoL) [3,4]. HRQoL (defined as ''physical, psychological, and social domains of health, seen as distinct areas that are influenced by a person's experiences, beliefs, expectations, and perceptions'') was also selected by an interdisciplinary group of stakeholders as a core outcome domain for clinical trials in LBP [5]. However, it is unclear which measurement instrument is best to measure this domain.

Conflict of interest: The authors of this manuscript declare that they do not have any conflict of interest related to the content of this manuscript.

Financial support: The authors of this manuscript would like to acknowledge the EUROSPINE task force research for providing funding for this study (EUROSPINE TFR 5-2015). This funding body did not have any role in designing the study, in collecting, analyzing and interpreting the data, in writing this manuscript, and in deciding to submit it for publication.

* Corresponding author: Department of Epidemiology and Biostatistics, Amsterdam Public Health research institute, Amsterdam Movement Sciences research institute, VU University Medical Center, de Boelelaan 1089a, Medical Faculty F-vleugel, 1081HV, Amsterdam, The Netherlands.

E-mail address: a.chiarotto@vumc.nl (A. Chiarotto).

https://doi.org/10.1016/j.jclinepi.2018.05.006
0895-4356/

What is new?

Key findings

- High quality evidence indicates that the construct validity of the summary scores of the SF-36 and of the utility and visual analogue scores of the EQ-5D-3L is inadequate in patients with low back pain (LBP).
- The quality of evidence on the content validity of six instruments widely used to measure HRQoL (i.e., SF-36, SF-12, EQ-5D-3L, EQ-5D-5L, NHP, PROMIS Global Health-10) is very low in patients with LBP.

What this adds to what is known?

- The measurement properties of HRQoL instruments have been only marginally investigated in patients with LBP.
- Caution should be used in assuming the validity of the SF-36 and EQ-5D-3L scores in patients with LBP, as there is high quality evidence suggesting that correlations of these scores with other instruments are not as expected.

What is the implication and what should change now?

- More research on the measurement properties of HRQoL instruments is needed in patients with LBP, and priority should be given to head-to-head comparison studies focusing on content validity.
- Future head-to-head comparisons should also assess structural validity, reliability, construct validity and responsiveness, and they should also include other recently developed instruments (e.g., LBP Core Set Self-Report Checklist, Musculoskeletal Health Questionnaire).

The selection of an instrument should be based on its measurement properties and feasibility in the target population [6,7]. Previous recommendations on HRQoL measurement in LBP have advocated the use of the Short Form 36 (SF-36), the Short Form 12 (SF-12) and/or the EuroQol 5D (EQ-5D) [8−12]. The SF-36 is most frequently used to measure HRQoL in LBP clinical trials, followed by the Nottingham Health Profile (NHP), the SF-12, the Sickness Impact Profile, and the EQ-5D [13]. The measurement properties of these instruments have been investigated in the general population and in various clinical samples [14]. However, it remains unclear how valid, reliable, and responsive these instruments are in patients with LBP.

Three reviews have attempted to summarize the measurement properties of HRQoL instruments in patients with LBP [15−17]. Two were narrative reviews [16,17], two focused on utility scores [15,17], and all had significant and important methodological weaknesses, such as failure to account for risk of bias in the evidence synthesis [18,19]. The Consensus-based Standards for the selection of health Measurement Instruments (COSMIN) initiative has developed tools to guide systematic reviews on measurement properties of patient-reported outcome measures [20]; these include a taxonomy defining each measurement property [21], a search filter to identify studies on measurement properties [22], a risk of bias assessment checklist [23], and evidence synthesis methods [24,25].

An international consortium developing a core outcome measurement set for LBP clinical trials selected five instruments as potential core outcome measurement instruments for HRQoL in LBP [26]. Four of these instruments (SF-36, SF-12, NHP and EQ-5D) were also among the five most frequently used in LBP trials; the Sickness Impact Profile was not selected because its length (136 items) rendered it unfeasible for inclusion in a core set [26]. Although it has not been broadly used, the 10-item Patient-Reported Outcome Measurement Information System (PROMIS) Global Health short form (PROMIS-GH-10) [27] was also chosen because it demonstrated face validity similar to the other instruments [26] and because it was recommended by another recent core set initiative [28].

This systematic review summarizes the evidence on the measurement properties of SF-36, SF-12, EQ-5D, NHP, and PROMIS-GH-10 in patients with LBP. The results of this review informed a Delphi survey to reach consensus on which instrument(s) to recommend for core outcome measurement of HRQoL in patients with LBP [26]. The original version of the EQ-5D includes three response options for each item (EQ-5D-3L) [29], and it has probably been the most used in LBP; however, because a version with five response options (EQ-5D-5L) was more recently developed [30], this newer EQ-5D version was also assessed in this review.

2. Methods

Conduct and report of this systematic review follows the COSMIN guidance [24] and the Preferred Reporting Items for Systematic Reviews and meta-Analysis statement [31]. The protocol was registered a priori in PROSPERO (https://www.crd.york.ac.uk/prospero/), number CRD42015020021.

2.1. Measurement instruments

2.1.1. Short Form 36

The SF-36 consists of 36 items measuring HRQoL subdivided in eight domains (Table 1). The number of response options varies from three (physical functioning subscale) to six (vitality and mental health subscales); originally, 0-100

scores were calculated for each domain, with higher scores indicating better health [32], but currently, these are usually expressed as t-scores. Two summary scores can be calculated: the physical component summary (PCS) and the mental component summary (MCS); these are calculated by summing factor-weighted scores across the eight subscales. For the PCS, highest weights are given to the physical functioning, role physical, bodily pain and general health scales; for the MCS, vitality, social functioning, role emotional, and mental health scales give higher weights. Factor weights are available from generic samples from various countries [14].

2.1.2. Short Form 12

The SF-12 was developed by selecting 12 items of the SF-36 to have a short HRQoL questionnaire (Table 1) [33]. It measures the same eight domains of the SF-36 and enables calculation of PCS and MCS summary scores. Similar to the SF-36, the SF-12 can also be obtained online [14].

2.1.3. 3-Level EuroQoL 5D

The EQ-5D-3L was developed as a HRQoL measure to be used in large-scale surveys and generate cross-national comparisons of health state valuations (Table 1) [29]. The first part of the instrument consists of five items measuring mobility, self-care, usual activities, pain/discomfort, and anxiety/depression with three response options each; these response options provide a profile (e.g., 21332) that is transformed into a utility score based on values from the general population [34]. The second part consists of a 0-100 vertical visual analogue scale (VAS) that scores self-rated health from "best imaginable health state" to "worst imaginable health state".

2.1.4. 5-Level EuroQol 5D

The utility part of the EQ-5D-3L was revised to improve the responsiveness and ceiling effects, by increasing the number of response options from three to five [30].

2.1.5. Nottingham Health Profile

The NHP was developed to capture perceived health status in the population (Table 1). It consists of 45 items with yes/no response options [35]. Thirty-eight items covering six domains are included in the first part that provides 0-100 scores for each domain; seven items covering seven different domains are included in the second part [14].

2.1.6. PROMIS global health short form

The PROMIS-GH-10 was developed to assess global health in patients with a variety of chronic conditions (Table 1). The items measure self-rated health, quality of life, physical functioning, psychological functioning, (satisfaction with) social roles and activities, fatigue, and pain [27]. All items are administered and scored on a 1-5 nominal scale, with the exception of the pain item that is administered on a 0-10 numeric rating scale and rescored to a 1-5 scale. Four items (two physical functioning, one fatigue, and one pain) are used to compute the PROMIS-GH-10 PCS score and four other items (two psychological functioning, one social satisfaction, and one quality of life) to compute the MCS score [27]; both scales are estimated using item response theory parameters. They are expressed as T-scores with a mean of 50 and standard deviation of 10, with higher scores indicating better health [36].

2.2. Literature search

2.2.1. Data sources and searches

MEDLINE (via PubMed), EMBASE (Embase.com), CINAHL (EBSCOhost), PsycINFO (EBSCOhost), and SportDiscus (EBSCOhost) were last searched on July 25, 2017. The search strategy consisted of three groups of search terms combined with the Boolean operator 'AND': (1) instrument names, (2) LBP, and (3) measurement properties. A previously developed search filter retrieved studies on the measurement properties in PubMed [22]; the same filter was adapted for all the other databases (Appendix 1). No restrictions for language or time were applied. Google Scholar was also searched (last on July 28, 2017) with the full names of the instruments, and the first 100 hits for each instrument were screened for inclusion. Citation tracking of the eligible studies was carried out in Web of Science and by hand searching reference lists.

2.2.2. Study selection

To select studies, COSMIN definitions for nine measurement properties were used [21]. Any report (e.g., book, online article) presenting the instrument development was included for the assessment of content validity [25]. Full-text of studies asking patients $\geq$18 years with nonspecific LBP [37] or professionals (e.g., researchers, clinicians) to assess relevance, comprehensiveness, and/or comprehensibility of an instrument were included as original content validity studies [25]. Studies that aimed to evaluate one of more of the other measurement properties were included if full-text articles presented the results for adult patients with nonspecific LBP. Studies that included patients with a mix of pathologies were included if at least 75% of the total sample had nonspecific LBP [38]. Studies that used the instruments only as outcome measurements, or for validation of other instruments, were excluded [24].

Inclusion criteria were applied by two reviewers (A.C. and S.J.K.) independently to titles and abstracts retrieved with the searches. Potentially eligible full-texts were also screened independently by two reviewers (A.C. and R.W.O.). Consensus on inclusion was sought between reviewers in a face-to-face meeting and, in case of disagreement, a third reviewer (C.B.T.) arbitrated.

2.3. Evaluation of the measurement properties

As per COSMIN guidance, measurement properties were assessed in the following order: (1) content validity,

Table 1. Characteristics and quality assessment of the studies on the development of the PROMs

PROM	Reference(s)	Language(s) (country) of development	Construct definition	Target population for that the PROM was developed	Intended context of use	Concept elicitation study		Cognitive interview study?	Overall development study quality
						Quality	Patients involved?		
SF-36	[32,60–63]	English (US)	"Health", eight concepts: physical functioning, social and role functioning, mental health, general health perceptions, bodily pain, and vitality.[a]	General population and patients	"Clinical practice and research, healthy policy evaluations, and general population surveys"	Inadequate	No	No	Inadequate
SF-12	[33]	English (US)	Not reported; assumed same of SF-36	Assumed same as SF-36	Assumed same as SF-36	Inadequate	No	No	Inadequate
EQ-5D-3L	[29,34]	Dutch English (UK) Finnish Norwegian Swedish	"Health-related quality of life"; no definition given	Large-scale surveys of the community and (…) for use in postal surveys	Complement other quality of life measures, collection of common data set for reference. Generate cross-national comparisons of health state valuations.	Inadequate	No	No	Inadequate
EQ-5D-5L	[30]	Spanish (Spain) English (UK)	Not reported; assumed same of EQ-5D-3L	Assumed same as EQ-5D-3L	Assumed same as EQ-5D-3L	Inadequate	Yes	Yes	Inadequate
NHP	[35,64,65]	English (UK)	Perceived health status in: physical mobility, pain, sleep, emotional reactions, social isolation, and energy.	"Population"	Identify groups in need of care, social policy (resource allocation), evaluate health and social services, identify consumer concerns, understand relationships between subjective responses to comparable pathologies.	Inadequate	No	No	Inadequate
PROMIS-GH-10	[27,66–69]	English (US)	"Global health": "person's general evaluation of health rather	National resource for precise and efficient PROMs on	Set of publicly available, efficient, and flexible PROMs,	Inadequate	No	No	Inadequate

(*Continued*)

Table 1. Continued

PROM	Reference(s)	Language(s) (country) of development	Construct definition	Target population for that the PROM was developed	Intended context of use	Concept elicitation study		Cognitive interview study?	Overall development study quality
						Quality	Patients involved?		
			than any of its specific components.'' Includes global ratings of the five primary PROMIS domains (physical function, fatigue, pain, emotional distress, and social health) and general health perceptions cutting across domains.	symptoms, functioning, and HRQoL, for patients with chronic diseases and conditions.	including HRQoL. Promoting the use of PROMs by the public and private sectors. (…). Repository of validated items and short forms and a CAT system.				

Abbreviations: EQ-5D-5L, 5-level 5-item EuroQol; EQ-5D-3L, 3-level 5-item EuroQol; NHP, Nottingham health profile; PROM, patient-reported outcome measure; PROMIS-GH-10, 10-item global health short form of the patient-reported outcomes measurement information system; SF36, short form 36; SF12, 12-item short form.

[a] Each of these health concepts had a separate definition that is not presented here because the scope of this review was to assess the SF36 as a measure of the construct HRQoL.

(2) internal structure (i.e., structural validity, internal consistency, and cross-cultural validity), (3) the remaining properties (i.e., test–retest reliability, measurement error, criterion validity, construct validity, responsiveness) [24].

2.3.1. Risk of bias assessment and data extraction

To assess content validity, the instrument development and original content validity studies should be evaluated [25]. The quality of these two different type of studies were assessed using newly developed standards (COSMIN risk of bias checklist) [23,25]. Each standard is rated on a four-point rating scale as ''very good,'' ''adequate,'' ''doubtful,'' or ''inadequate''. Total scores were determined for two parts of the development study (concept elicitation and cognitive interview) separately. Furthermore, each aspect of a content validity study (i.e., relevance, comprehensiveness, comprehensibility) was evaluated separately. A total score is obtained for each part by taking the lowest rating among the standards (i.e., worst-score counts) [39]. More detailed information on these new standards for content validity can be found elsewhere [25]. Two reviewers (A.C. and C.B.T.) assessed studies examining this measurement property separately and reached consensus in a face-to-face meeting.

The risk of bias of the studies on the other measurement properties was also assessed with the COSMIN risk of bias checklist [23]. The four-point rating scale and worst-score counts method are the same for every measurement property, and a total score is provided for studies of each measurement property in each study. Two reviewers (A.C. and R.W.O.) assessed the risk of bias independently and reached consensus in a face-to-face meeting.

For every study, data on patients' characteristics and results were extracted by one reviewer (A.C.), and a random 25% of the extracted information, stratified per instrument, was double checked by a second reviewer (R.W.O.).

2.3.2. Evidence synthesis

For content validity, the results of each study were rated by two reviewers (A.C. and C.B.T.) independently against 10 established criteria: five on relevance, one on comprehensiveness, and four on comprehensibility [25]. Each criterion could be rated as sufficient (+), insufficient (−), or indeterminate (?). The same criteria were also scored based on the content of the instrument itself, and reviewers found consensus in a face-to-face meeting [25]. Subsequently, the results of all studies on a specific instrument and the reviewer's rating were summarized qualitatively. An overall sufficient (+), insufficient (−), or inconsistent (±) rating was provided for relevance, comprehensiveness, and comprehensibility of each instrument.

For the other measurement properties, the results were rated according to the consensus-based criteria proposed by Prinsen et al. [7]. A priori hypotheses were formulated by the review team to evaluate the results of studies on construct validity and responsiveness, based on the results of a previous systematic review on instruments for LBP [40]. For both properties, correlations were expected to be:

- ≥ 0.60 with instruments measuring the same construct (e.g., physical component summary [PCS] or physical functioning subscale with a physical functioning instrument or subscale);
- < 0.60 and ≥0.30 with instruments measuring largely related but dissimilar constructs (e.g., PCS or physical functioning subscale with a pain instrument or subscale);
- < 0.50 and ≥0.20 with instruments measuring moderately related but dissimilar constructs (e.g., PCS or physical functioning subscale with a mental health instrument or subscale);
- < 0.30 with instruments measuring weakly related or unrelated constructs (e.g., PCS of physical functioning subscale with a general health instrument or subscale).

Two additional hypotheses were formulated for responsiveness:

- The area under the curve to discriminate between improved and unchanged patients (as defined in each study) had to be ≥0.70;
- Effect size and standardized response means for improved patients (as defined in each study) had to be at least 0.50 larger than those for unchanged patients [41].

Within studies, construct validity and responsiveness were considered sufficient (+) if ≥75% of the hypotheses were met, or insufficient (−) if ≥75% of the hypotheses were not met, otherwise they were considered inconsistent (±) [7]. Between studies, results were considered inconsistent if they did not display the same results (i.e., all sufficient, insufficient or inconsistent findings).

The quality of evidence for each measurement property was rated according to the Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) approach [42] adapted for this type of review into: "high," "moderate," "low," or "very low" [24,25]. For content validity, the evidence quality could be downgraded because of risk of bias, inconsistency of results and indirectness, as outlined elsewhere [25]. For the other measurement properties, risk of bias, imprecision, inconsistency, and indirectness were taken into account [24]. Quality of evidence was downgraded as follows:

- Risk of bias: one level if there was only one "adequate" study, two levels if there was only one or more "doubtful" studies, and three levels if there was only one or more "inadequate" studies;
- Imprecision: one level if the total patient sample was <100 and ≥50, two levels if <50;
- Inconsistency: one level if the studies displayed inconsistent findings (e.g., one study sufficient, another study insufficient results);
- Indirectness: one level if a study did not specifically address the construct (HRQoL) and/or the target population (adult patients with nonspecific LBP) [43].

3. Results

From 4,809 records identified in the electronic searches, 95 full-text articles were retrieved, and 26 finally deemed eligible; a further nine eligible studies were identified by citation tracking; therefore, 35 studies were included (Fig. 1). Sixty-nine full-texts were excluded: 14 did not include patients with LBP, 13 did not assess a measurement property as defined by COSMIN, 11 did not present results separately for patients with nonspecific LBP, 10 included patients with specific LBP, seven focused on measurement properties of other instruments, five were unclear if they included patients with nonspecific LBP, five analyzed only a part of the eligible instruments (e.g., physical functioning subscale of the SF-36), two were not relevant for this review, one was a review, and one assessed an instrument that was not self-reported but administered by an interviewer.

Seventeen of the included articles presented information on the development of the instruments (Table 1). Among the other 18 included studies, three evaluated the SF-36 and the EQ-5D-3L simultaneously [44–46], six only the SF-36 [47–52], three only the SF-12 [53–55], and four only the EQ-5D-3L [56–59]. The characteristics of these studies and of the patients included are displayed in Table 2.

3.1. Short Form 36

Two books and three articles [32,60–63] reported information on the SF-36 development (Table 1). The other nine articles assessed structural validity, internal consistency, test–retest reliability, construct validity, and/or responsiveness (Table 2). Eight studies reported on the SF-36 subscale scores, three studies on PCS and MCS scores, and one evaluated the total score (Table 2).

3.1.1. Content validity

The SF-36 development was considered of inadequate quality because no patients were involved, and no cognitive interview study was performed (Table 1). Because no additional content validity studies were found, solely the reviewers' ratings counted for the evidence synthesis, leading to very low quality evidence of sufficient content validity (Table 3).

3.1.2. Internal structure

Zwingmann et al. [52] assessed SF-36 structural validity in a study of doubtful quality. Principal component analysis with varimax rotation and confirmatory factor analysis were applied at the subscale level. Component analysis extracted two factors with eigenvalues equal to 2.3 and 2.8, respectively (total explained variance 63%). Three subscales (physical functioning, role physical, and bodily pain) loaded >0.70 on the PCS, the other five >0.50 on the MCS. Confirmatory analysis found the same subscales loading on the same components (Goodness-of-Fit index = 0.96, adjusted Goodness-of-Fit index = 0.91), with the exception of the general health subscale that loaded about 0.30 on both

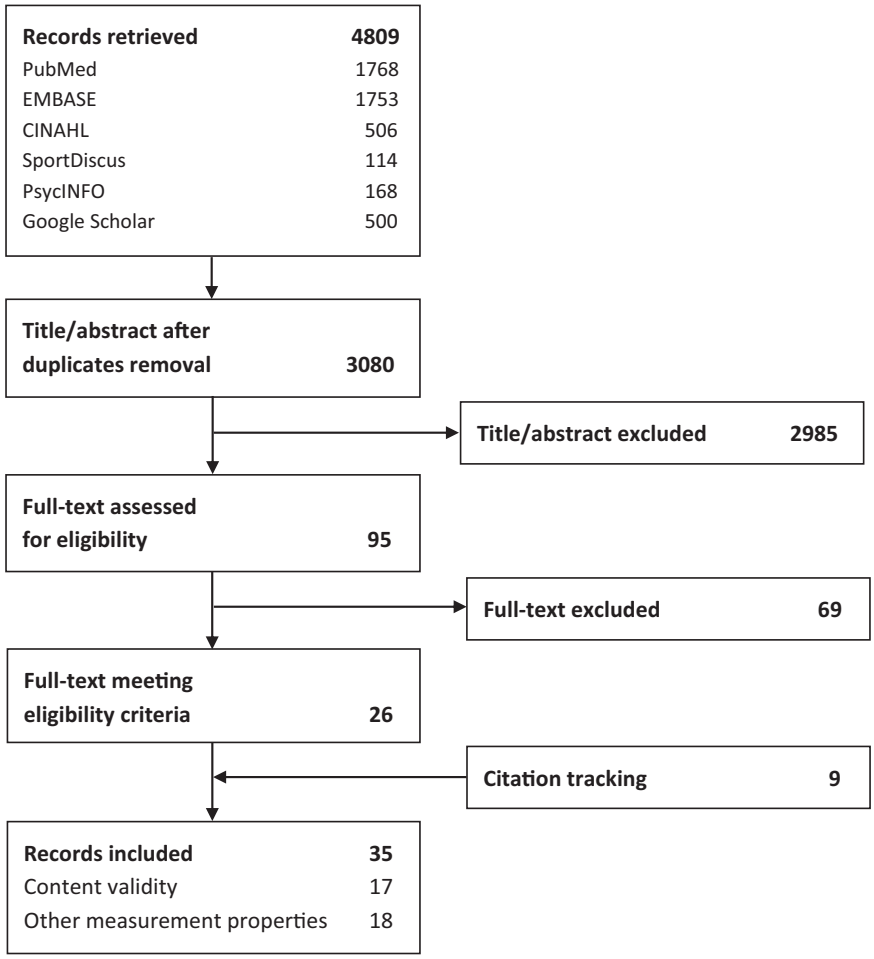

Fig. 1. Flow chart of results of search strategy and selection of records.

PCS and MCS. Considering that the general health subscale is supposed to load on the MCS, these results on the unidimensionality of the two component scores were considered inconsistent based on low quality evidence.

Two studies [48,52] of very good quality assessed the internal consistency of the subscales, finding a Cronbach's alpha >0.70 for seven of eight subscales in each study. According to recent COSMIN guidance [24], because no studies assessed the unidimensionality of the SF-36 subscales, these results cannot provide evidence in support of this measurement property. One study [47] evaluated internal consistency of the total scale in a study of inadequate quality (Cronbach's alpha = 0.80). No studies assessed cross-cultural validity.

3.1.3. Other measurement properties

In one study of doubtful quality (Table 4), sufficient test–retest reliability was found for the subscales (low quality evidence). No studies assessed test–retest reliability of PCS, MCS, and total score; likewise, measurement error was not evaluated.

Construct validity was evaluated for subscales, PCS, MCS, and total scores (Appendix 2). Low quality evidence (due to risk of bias and inconsistent results across studies) of inconsistent construct validity was found for the subscales (Table 3). High quality evidence of insufficient construct validity was found for both PCS and MCS scores (Table 3). For the total score, there was low quality evidence (due to risk of bias) of sufficient construct validity (Table 3).

Responsiveness was also evaluated for subscales, PCS, MCS, and total scores (Appendix 3). For the subscales, hypotheses could not be tested in three studies [50–52] because only overall effect sizes were reported; in the study by Suarez-Almazor et al. [46], inconsistent results were found (very low quality evidence due to risk of bias and imprecision) as 11 of 32 hypotheses were met (34%) (Table 3). Low quality evidence (due to risk of bias and inconsistency) of inconsistent responsiveness was found for PCS and MCS scores (Table 3). Sufficient responsiveness was found for the SF-36 total score (low quality evidence, Table 3), based on one study [47] with an area under the curve >0.70.

3.2. Short Form 12

One study [33] reported information on the SF-12 development; three studies [53–55] assessed the following

Table 2. Characteristics of the studies assessing the measurement properties of HRQoL instruments in patients with low back pain

PROM(s)	Reference	Language (country)	Study design	Health intervention(s)	LBP specifics
SF-36; EQ-5D-3L	Suarez-Almazor 2000 [46]	English (Canada)	Longitudinal		>3 mo
SF-36; EQ-5D-3L	Campbell 2006 [44]	English (UK)	RCT	Surgical spine stabilization or intensive rehabilitation	>12 mo
SF-36; EQ-5D-3L	Eker 2007 [45]	Turkish	Cross-sectional		>3 mo
SF-36	Bullinger 1995 [48]	German (Germany)	Longitudinal		
SF-36	Gatchel 1998 [50]	English (US)	Longitudinal	Functional restoration management	>4 mo
SF 36	Zwingmann 1998 [52]	German (Germany)	Cross-sectional	Inpatient rehabilitation	
SF-36	Bronfort 1999 [47]	English (US)	RCT	Exercise vs exercise plus manual therapy or NSAIDs	Chronic (not defined) ± leg pain
SF-36	Dunn 2003 [49]	English (UK)	Longitudinal		
SF-36	Weigl 2006 [51]	German (Germany)	Longitudinal	Individual health resort programs	
SF-12	Luo 2003 [54]	English (US)	Longitudinal		
SF-12	Turner 2003 [55]	English (US)	Longitudinal		Work-related claim
SF-12	Diaz-Arribas 2017 [53]	Spanish (Spain)	Longitudinal	Conservative management	≥14 d
EQ-5D-3L	Garratt 2001 [56]	English (UK)	RCT	Exercise with cognitive behavioral approach or usual GP care	>4 wk and <6 mo
EQ-5D-3L	Mueller 2013 [57]	English (US)	Cross-sectional		
EQ-5D-3L	Soer 2012 [58]	Dutch	Longitudinal	Multidisciplinary rehabilitation or anesthesiology	>3 mo
EQ-5D-3L	Whynes 2013 [59]	English (UK)	RCT	Epidural steroid injections	± leg pain

Abbreviations: EQ-5D-3L, 3-level EuroQol 5D; *n*, sample size; NSAIDs, non-steroidal anti-inflammatory drugs; PROM(s), patient-reported outcome measure(s); RCT, randomized controlled trial; SD, standard deviation; SF-12, short form 12 health survey; SF-36, short form 36 health survey; VAS, visual analogue scale; μ, mean.
Empty cells indicate not assessed items.
[a] Measurement error was assessed in 163 patients, responsiveness in 359, but characteristics of these patients alone were not presented.

measurement properties: internal consistency, measurement error, construct validity, and responsiveness (Table 2). Each study evaluated SF-12 PCS and MCS scores.

3.2.1. Content validity

The SF-12 was derived from the SF-36 using regression methods [33]; its development was considered inadequate as no patients were involved, and no cognitive interview study was performed (Table 1). No additional content validity studies were retrieved. There was very low quality evidence of sufficient content validity (based on the reviewers' ratings only) (Table 3).

3.2.2. Internal structure

No studies evaluated the SF-12 structural and cross-cultural validity. The only study [54] assessing internal

Measurement properties	Scores used	Patient characteristics				
		n	Female (%)	Age (μ ± SD, years)	Pain duration (μ ± SD, years)	Working (%)
Responsiveness	Subscales and component summaries (SF-36). Utility and VAS (EQ-5D)	46	65	50 ± 15	10 ± 11	
Responsiveness	Component summaries (SF-36). Utility (EQ-5D)	250	44	40 ± 9	8 ± 7	44
Construct validity	Subscales (SF-36). Domains (EQ-5D)	132	69	59 ± 13	10 ± 12	
Internal consistency, construct validity	Subscales	144				
Construct validity, responsiveness	Subscales and component summaries	188	40	41 ± 10		0
Structural validity, internal consistency, responsiveness	Subscales	244	26	44 ± 8		
Internal consistency, construct validity, responsiveness	Total (all properties) and subscales (construct validity)	132	51	42 ± 9	2 (median)	92 (full-time)
Test–retest reliability	Subscales	14	57	45 ± 8		
Construct validity, responsiveness	Subscales	178	52	66 ± 8		
Internal consistency, construct validity, responsiveness	Component summaries	2520	55	52 ± 16		31
Construct validity, responsiveness	Component summaries	309	37	41 ± 11		76
Measurement error, responsiveness	Component summaries	458[a]	78	46 ± 11	1 ± 1	82
Construct validity, responsiveness	Utility	179	57			
Construct validity	Utility, VAS and domains	8,385	60	53 ± 16		
Construct validity, responsiveness	Utility and VAS	151	55	52 ± 16		47
Responsiveness	Utility and VAS	39				

consistency was of very good quality and found Cronbach's alpha equal to 0.77 and 0.80 for PCS and MCS scores, respectively. This study does not represent evidence on this measurement property because the unidimensionality of these scores was not assessed in any study [24].

3.2.3. *Other measurement properties*

No studies assessed SF-12 test–retest reliability. Diaz-Arribas et al. [53] (doubtful quality) was the only study assessing measurement error. The MIC values presented in this study were larger than the smallest detectable change values (Table 4), providing low quality evidence for sufficient measurement error for both PCS and MCS.

Moderate quality evidence (downgraded for risk of bias) (Appendix 2) of sufficient construct validity was found for PCS, although low quality evidence (risk of bias and inconsistent results) of inconsistent construct validity for MCS (Table 3). Very low quality evidence (due to risk of bias

Table 3. Evidence synthesis on measurement properties of HRQoL instruments in patients with low back pain

	SF-36				SF-12		EQ-5D-3L			EQ-5D-5L			NHP	PROMIS-GH-10	
Measurement properties	**Subscales**	**PCS**	**MCS**	**TOT**	**PCS**	**MCS**	**Utility**	**VAS**	**Domains**	**Utility**	**VAS**	**Domain**	**Subscales**	**PCS**	**MCS**
Content validity															
Relevance															
Rating	+				+		+			+			+		
Quality	All: very low														
Comprehensiveness															
Rating	+				–		–			+			+		
Quality	All: very low														
Comprehensibility															
Rating	+				+		+			+			+		
Quality	All: very low														
Construct validity															
Rating	±	–	–	+	+	±	–	–	±				+		
Quality	Low	High	High	Low	Modera-te	Low	High	High	Low				Low		
Responsiveness															
Rating	±	±	±	+	±	–	±	±							
Quality	Very low	Low	Low	Low	Very low	Low	Moderate	Low							

Abbreviations: EQ-5D-3L, 3-level EuroQoL 5D; EQ-5D-5L, 5-level EuroQoL 5D; MCS, mental component summary; NHP, Nottingham health profile; PCS, physical component summary; PROMIS-GH-10, 10-item PROMIS global health short form; SF-36, short form 36; SF-12, short form 12; VAS, visual analogue scale; "+", sufficient results; "–", insufficient results; "±", conflicting results.

Empty cells represent measurement properties not assessed in any study.

The following measurement properties were assessed only for one instrument, and their evidence synthesis is not reported in the table: structural validity of SF-36 PCS and MCS was inconsistent (low quality evidence), test–retest reliability of SF-36 subscales was sufficient (low quality evidence), measurement error of SF-12 PCS and MCS was sufficient (low quality evidence). Internal consistency was assessed for the SF-36 subscales, SF-36 TOT, and for SF-12 PCS, and MCS; the results of these studies are presented in the text, but they do not represent for this measurement property, as the unidimensionality of the assessed tools has not been tested in patients with low back pain. Cross-cultural validity was not assessed for any instrument.

and inconsistency) underpinned inconsistent responsiveness of the PCS, although low quality evidence (because of risk of bias) was present for insufficient responsiveness of the MCS (Table 3).

3.3. 3-Level EuroQoL 5D

Two studies reported information on the EQ-5D-3L development [29,34]; other seven articles [44–46,56–59] assessed its measurement properties, six focusing on the utility score, four on the VAS, and two on the domain scores (Table 2).

3.3.1. Content validity

The EQ-5D-3L development was rated inadequate because no patients were involved in the instrument development, no clear definition of the construct to be measured was provided, and no cognitive interview study was performed (Table 1). There was very low quality evidence of sufficient relevance and comprehensibility and very low quality evidence of insufficient comprehensiveness (Table 3).

3.3.2. Internal structure

No studies tested structural validity, internal consistency, or cross-cultural validity in patients with LBP.

3.3.3. Other measurement properties

Test–retest reliability and measurement error were not assessed in any study. High quality evidence for insufficient construct validity (Appendix 2) was found for utility and VAS scores (Table 3). For the domain scores, low quality evidence (due to risk of bias) was found for inconsistent results (Table 3) because two studies of doubtful quality reported inconsistent results [45,57].

There was moderate quality evidence for inconsistent responsiveness of the utility score (Table 3), as results were largely inconsistent across studies (Appendix 3). Low quality evidence (because of risk of bias and inconsistency) was found for inconsistent responsiveness of the VAS score (Table 3). No studies assessed the responsiveness of the domain scores.

3.4. 5-Level EuroQol 5D

One article reported information on the development of the EQ-5D-5L [30] and was rated as inadequate because a clear description of the construct to be measured was not provided as it referred to the development of the EQ-5D-3L (Table 1). A pilot study (focus groups) was conducted to assess the comprehensibility (but not comprehensiveness) of the instrument; that study was rated as doubtful quality because it was

Table 4. Test–retest reliability and measurement error of HRQoL instruments in patients with low back pain

PROM	Reference	n	Study quality	Time interval	Scores used	Test–retest reliability ICC (lower limit, 95% CI)
SF-36	Dunn 2003 [49]	14	Doubtful	2 wk	Subscale:	
					Physical functioning	0.93 (0.84)
					Role physical	0.81 (0.51)
					Role emotional	0.74 (0.43)
					Social functioning	0.88 (0.71)
					Mental health	0.90 (0.76)
					Vitality	0.94 (0.85)
					Bodily pain	0.89 (0.63)
					General health	0.96 (0.90)
						Measurement error
						SDC (% scale range)
SF-12	Diaz-Arribas 2017 [53]	163	Doubtful	12 mo	Component summary:	
					Physical	0.56 (1)[a]
					Mental	3.77 (6)[a]

Abbreviations: 95% CI, 95% confidence interval; ICC, intraclass correlation coefficient; *n*, number; PROM, patient-reported outcome measure; SDC, smallest detectable change.

[a] Unclear that SDC formula was used.

unclear if the content analyses were performed by two assessors. The reviewers rated the EQ-5D-5L content validity for patients with LBP as sufficient for relevance and comprehensibility (very low quality evidence), and as insufficient for comprehensiveness (very low quality) (Table 3).

No studies assessed the measurement properties of this instrument in patients with LBP.

3.5. Nottingham Health Profile

Three studies reported the development of the NHP [35,64,65], which was rated as inadequate because patients were not involved in the concept elicitation study (Table 1). No studies assessed the content validity. There was very low quality evidence for sufficient content validity, as rated by the reviewers (Table 3).

No studies were found on the internal structure, reliability, and/or responsiveness of the NHP in patients with LBP. One study [48] assessed the construct validity of its subscales as compared to the SF-36 subscales, providing low quality evidence (due to risk of bias, Table 3) of sufficient construct validity as 37 of 48 hypotheses were met (77%, Appendix 2).

3.6. PROMIS global health short form

Four articles and one report included information on the PROMIS-GH-10 development [27,66–69]. Development was rated inadequate because, despite patients' involvement in developing PROMIS items [69], they were not specifically involved in selecting the most relevant (sub)domains to be included in this instrument (Table 1). There was very low quality evidence of sufficient content validity based solely on the reviewers' ratings (Table 3). No studies assessing the measurement properties of the PROMIS-GH-10 in patients with LBP were retrieved.

4. Discussion

This systematic review highlights the scarcity of high quality evidence on the measurement properties of SF-36, SF-12, EQ-5D-3L, EQ-5D-5L, NHP, and PROMIS-GH-10 in patients with LBP. Foremost, there is very low quality evidence for the content validity of these six instruments in patients with LBP (Table 3). There was high quality evidence only of insufficient construct validity of SF-36 physical and mental summary scores, and EQ-5D-3L utility and VAS scores. Construct validity and responsiveness were the most often assessed properties, mainly for SF-36, SF-12, and EQ-5D-3L. Moderate quality evidence was found for SF-12 PCS score inconsistent construct validity and EQ-5D utility score inconsistent responsiveness (Table 3). All other properties of any instrument were underpinned by lower quality evidence or not assessed.

Content validity is considered by clinimetric and psychometric experts as the first measurement property to consider when selecting an instrument [7,25]. The six instruments assessed in this review were developed without patient input (Table 1). For this reason, content validity of these instruments should be urgently assessed in patients with LBP. The HRQoL definition used in a recent consensus exercise on core outcome domains for LBP

may be adopted as a starting point [5]. This definition highlights three major subdomains (i.e., physical, mental, and social) but does not provide much detail on what should be included within each subdomain. It should be recognized that there is no consensus on the definition of HRQoL and that other definitions exist [70−72]. Patients have indicated that an acceptable level of HRQoL is required for recovery from an LBP episode [4], but we did not explore what HRQoL really means for them. This issue also needs to be addressed.

The structural validity (second measurement property to consider when selecting an instrument [7]) of HRQoL instruments in patients with LBP has not been sufficiently investigated for the SF-36 (Table 3). The EQ-5D instruments provide two ''unidimensional'' HRQoL scores (utility index and VAS), the other instruments provide component summary and/or subscale scores. It is unclear which performs best in patients with LBP. HRQoL definitions usually highlight the multidimensional nature of the domain, but it remains unclear if HRQoL can be adequately expressed using the total score of a multidimensional instrument or only by the subscales scores. Head-to-head comparison of available instruments is required, and various psychometric methods, such as (exploratory and confirmatory) factor analysis, bifactor analysis, Mokken scale analysis, or parametric item response theory, are available to address this issue [7]. An alternative may be to consider the total scores of these instruments as formative models, and this would not require assessment of structural validity. However, to consider these instruments as based on formative HRQoL models, there should be high quality evidence on their sufficient content validity. Scores from the same HRQoL instruments have been pooled in systematic reviews on interventions [38,73], under the assumption that validity is the same across cultures, countries and languages; however, because there are no studies on the cross-cultural validity of these instruments (Table 3), differential item functioning should be tested to see if this cross-cultural equivalence assumption holds [74].

The most important finding regarding the other measurement properties assessed in this review is the insufficient construct validity of the SF-36 PCS and MCS scores and of the EQ-5D-3L utility and VAS scores. Based on these results, it is essential to better assess the construct validity of SF-36 subscales as these may display better results than summary scores. Regarding the EQ-5D, it should be underlined that only the EQ-5D-3L version has been assessed in patients with LBP and not the EQ-5D-5L [30]. This newer version has shown improved validity in other populations [75]. Some researchers do not consider the EQ-5D utility score a patient-reported outcome measure of HRQoL [26] because it is not based on direct patient-reporting, but rather on values derived from the general population. Others disagree, equating the valuation of EQ-5D profiles to the use of norm scores, as applied in PROMIS-GH-10.

This review included six instruments among those most frequently used or recommended for HRQoL [8−13,28]; however, other instruments should also be included in future head-to-head comparisons in patients with LBP. Bagraith et al. [76,77] have recently developed the LBP Core Set Self-Report Checklist that was based on the categories included in the ICF core set for LBP [78]. This unidimensional tool includes both activity limitations and participation restrictions, it could be considered as an HRQoL instrument and compared to the others included in this review. The same argument holds for another recently developed tool, the Musculoskeletal Health Questionnaire [79], which aims at including all domains relevant to patients with musculoskeletal complaints.

The results of this systematic review informed an international and multidisciplinary Delphi process to endorse core instruments for LBP clinical trials; however, none of the HRQoL instruments met the prespecified 67% cutoff for consensus (although the SF-12 came close) [26]. To improve measurement standardization, the international consortium overseeing the core set initiative for LBP decided to recommend the SF-12 and the PROMIS-GH-10 for HRQoL, as they both provide PCS and MCS scores, and the PROMIS-GH-10 represents a free of charge option [26]. On the other hand, the SF-36 was not endorsed because it was considered too long for inclusion in every clinical trial, and the EQ-5D because it provided scores that cannot be pooled with the other instruments' scores (i.e., SF12 and PROMIS-GH-10) [26]. However, these recommendations are provisional, and taken together with the results of this review, reinforce the message that substantial input is needed from the scientific community to generate more evidence on HRQoL measurement in LBP.

This study was conducted according to current guidance on systematic reviews of patient-reported outcome measures [7,21,23−25]. However, the reviewed instruments were developed before the more recent developments in clinimetrics and psychometrics. Specifically, they could not consult the COSMIN risk of bias checklist [23] in the design phase of the study. Additionally, given that the EQ-5D-5L and PROMIS-GH-10 were developed 10 to 20 years later than the other instruments, it is no surprise that studies in patients with LBP are lacking. This is a reason why there are less studies assessing the measurement properties of these instruments. We reiterate that these newer instruments should be assessed in head-to-head comparisons in patients with LBP.

The HRQoL instruments assessed in this review have been assessed in several other populations [14], but, up to now, there is no evidence showing whether it is appropriate to translate findings on measurement properties from generic populations to populations with specific health conditions and between health conditions. In this review, we focused on their measurement properties in patients with LBP following the approach routinely used for systematic reviews of clinical trials, observational and

diagnostic studies in the LBP field [38,73,80−85]. A potential limitation is that the evidence syntheses were performed lumping together studies from different cultures, countries, and languages. The same approach is routinely used in systematic reviews of clinical trials and observational studies in patients with back pain disorders [38,73,80,82−86]; splitting the results of these studies may be considered equally contentious as there is no evidence clearly indicating the best way to synthesize evidence on instrument measurement properties. Meanwhile, for more detailed scrutiny, we provided quality and results of each study specifying language and country (Tables 2 and 4, Appendix 2 and 3). Another potential limitation of this review is that the search strategy was designed to identify studies on patients with LBP (Appendix 1), without specifically targeting the development studies of the instruments that were conducted in generic populations; nevertheless, eight articles on the instruments development were found with the initial search and the remaining nine through citation tracking (Figure 1).

Evidence related to the measurement properties of HRQoL instruments in patients with LBP is inconclusive (Table 3). With a few exceptions, higher quality evidence is needed on all measurement properties of the instruments included in this review and other instruments. Head-to-head comparisons of the content validity of the various instruments should have priority.

Acknowledgments

None.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2018.05.006.

References

[1] GBD 2016 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990-2016: a systematic analysis for the global burden of disease study 2016. Lancet 2017;390:1211−59.

[2] Maniadakis N, Gray A. The economic burden of back pain in the UK. Pain 2000;84:95−103.

[3] Froud R, Patterson S, Eldridge S, Seale C, Pincus T, Rajendran D, et al. A systematic review and meta-synthesis of the impact of low back pain on people's lives. BMC Musculoskelet Disord 2014;15:50.

[4] Hush JM, Refshauge K, Sullivan G, De Souza L, Maher CG, McAuley JH. Recovery: what does this mean to patients with low back pain? Arthritis Rheum 2009;61:124−31.

[5] Chiarotto A, Deyo RA, Terwee CB, Boers M, Buchbinder R, Corbin TP, et al. Core outcome domains for clinical trials in non-specific low back pain. Eur Spine J 2015;24:1127−42.

[6] Boers M, Brooks P, Strand CV, Tugwell P. The OMERACT filter for outcome measures in rheumatology. J Rheumatol 1998;25:198−9.

[7] Prinsen CA, Vohra S, Rose MR, Boers M, Tugwell P, Clarke M, et al. How to select outcome measurement instruments for outcomes included in a "Core Outcome Set"−a practical guideline. Trials 2016;17:449.

[8] Bombardier C. Outcome assessments in the evaluation of treatment of spinal disorders: summary and general recommendations. Spine 2000;25:3100−3.

[9] Chiarotto A, Terwee CB, Ostelo RW. Choosing the right outcome measurement instruments for patients with low back pain. Best Pract Res Clin Rheumatol 2016;30:1003−20.

[10] Clement RC, Welander A, Stowell C, Cha TD, Chen JL, Davies M, et al. A proposed set of metrics for standardized outcome reporting in the management of low back pain. Acta Orthop 2015;86:523−33.

[11] Deyo RA, Battie M, Beurskens A, Bombardier C, Croft P, Koes B, et al. Outcome measures for low back pain research: a proposal for standardized use. Spine 1998;23:2003−13.

[12] Resnik L, Dobrzykowski E. Guide to outcomes measurement for patients with low back pain syndromes. J Orthop Sports Phys Ther 2003;33:307−16.

[13] Chapman JR, Norvell DC, Hermsmeyer JT, Bransford RJ, DeVine J, McGirt MJ, et al. Evaluating common outcomes for measuring treatment success for chronic low back pain. Spine 2011;36:S54−68.

[14] Busija L, Pausenberger E, Haines TP, Haymes S, Buchbinder R, Osborne RH. Adult measures of general health and health-related quality of life: medical outcomes study short form 36-item (SF-36) and short form 12-item (SF-12) health surveys, Nottingham health profile (NHP), sickness impact profile (SIP), medical outcomes study short form 6D (SF-6D), health utilities index mark 3 (HUI3), quality of well-being scale (QWB), and assessment of quality of life (AQoL). Arthritis Care Res (Hoboken) 2011;63(Suppl 11):S383−412.

[15] Finch AP, Dritsaki M, Jommi C. Generic preference-based measures for low back pain: which of them should be used? Spine 2016;41:E364.

[16] Lurie J. A review of generic health status measures in patients with low back pain. Spine 2000;25:3125−9.

[17] Tosteson AN. Preference-based health outcome measures in low back pain. Spine 2000;25:3161−6.

[18] Higgins JP, Green S. Cochrane handbook for systematic reviews of interventions. John Wiley & Sons; 2011. Available at: handbook-5-1.cochrane.org.

[19] Katikireddi SV, Egan M, Petticrew M. How do systematic reviews incorporate risk of bias assessments into the synthesis of evidence? A methodological study. J Epidemiol Community Health 2015;69:189−95.

[20] Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, et al. Protocol of the COSMIN study: COnsensus-based Standards for the selection of health Measurement INstruments. BMC Med Res Methodol 2006;6:2.

[21] Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol 2010;63:737−45.

[22] Terwee CB, Jansma EP, Riphagen II, de Vet HC. Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. Qual Life Res 2009;18:1115−23.

[23] Mokkink LB, De Vet HC, Prinsen CA, Patrick DL, Alonso J, Bouter LM, et al. COSMIN risk of bias checklist for systematic reviews of patient-reported outcome measures. Qual Life Res 2018;27:1171−9.

[24] Prinsen CA, Mokkink LB, Bouter LM, Alonso J, Patrick DL, De Vet HC, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. Qual Life Res 2018;27:1147−57.

[25] Terwee CB, Prinsen CA, Chiarotto A, De Vet HC, Westerman MJ, Patrick DL, et al. COSMIN methodology for evaluating the content

validity of patient-reported outcome measures: a Delphi study. Qual Life Res 2018;27:1159−70.
[26] Chiarotto A, Boers M, Deyo RA, Buchbinder R, Corbin TP, Costa LO, et al. Core outcome measurement instruments for clinical trials in non-specific low back pain. Pain 2018;159:481−95.
[27] Hays RD, Bjorner JB, Revicki DA, Spritzer KL, Cella D. Development of physical and mental health summary scores from the patient-reported outcomes measurement information system (PROMIS) global items. Qual Life Res 2009;18:873−80.
[28] Salinas J, Sprinkhuizen SM, Ackerson T, Bernhardt J, Davie C, George MG, et al. An international standard set of patient-centered outcome measures after stroke. Stroke 2016;47:180−6.
[29] Group TE. EuroQol-a new facility for the measurement of health-related quality of life. Health Policy 1990;16:199−208.
[30] Herdman M, Gudex C, Lloyd A, Janssen M, Kind P, Parkin D, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual Life Res 2011;20:1727−36.
[31] Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009;6:e1000097.
[32] Ware JE Jr, Sherbourne CD. The MOS 36-item short-form health survey (SF-36): I. Conceptual framework and item selection. Med Care 1992;30:473−83.
[33] Ware JE Jr, Kosinski M, Keller SD. A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. Med Care 1996;34:220−33.
[34] Brooks R, Group E. EuroQol: the current state of play. Health Policy 1996;37:53−72.
[35] Hunt SM, McEwen J. The development of a subjective health indicator. Sociol Health Illn 1980;2:231−46.
[36] Center^{SM} HSSpbA. An application to score PROMIS® and Neuro-QoL instruments. Available at: https://www.assessmentcenter.net/ac_scoringservice2017. Accessed November 14, 2017.
[37] Maher C, Underwood M, Buchbinder R. Non-specific low back pain. Lancet 2017;389:736−47.
[38] Kamper SJ, Apeldoorn AT, Chiarotto A, Smeets RJ, Ostelo RW, Guzman J, et al. Multidisciplinary biopsychosocial rehabilitation for chronic low back pain. Cochrane Database Syst Rev 2014;Cd000963.
[39] Terwee CB, Mokkink LB, Knol DL, Ostelo RW, Bouter LM, de Vet HC. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Qual Life Res 2012;21:651−7.
[40] Chiarotto A, Maxwell LJ, Terwee CB, Wells GA, Tugwell P, Ostelo RW. Roland-Morris disability questionnaire and Oswestry Disability Index: which has better measurement properties for measuring physical functioning in nonspecific low back pain? Systematic review and meta-analysis. Phys Ther 2016;96:1620−37.
[41] De Vet HC, Terwee CB, Mokkink LB, Knol DL. Measurement in medicine: a practical guide. New York, USA: Cambridge University Press; 2011.
[42] Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;336:924.
[43] Chiarotto A, Ostelo RW, Boers M, Terwee CB. A systematic review highlights the need to investigate the content validity of patient-reported outcome measures for physical functioning in low back pain. J Clin Epidemiol 2018;95:73−93.
[44] Campbell H, Rivero-Arias O, Johnston K, Gray A, Fairbank J, Frost H. Responsiveness of objective, disease-specific, and generic outcome measures in patients with chronic low back pain: an assessment for improving, stable, and deteriorating patients. Spine (Phila Pa 1976) 2006;31:815−22.
[45] Eker L, Tuzun E, Daskapan A, Bastug Z, Yakut Y. The relationship between EQ-5D and SF-36 instruments in patients with low back pain. Fizyoterapi Rehabilitasyon 2007;18:3.
[46] Suarez-Almazor ME, Kendall C, Johnson JA, Skeith K, Vincent D. Use of health status measures in patients with low back pain in clinical settings. Comparison of specific, generic and preference-based instruments. Rheumatology (Oxford) 2000;39:783−90.
[47] Bronfort G, Bouter LM. Responsiveness of general health status in chronic low back pain: a comparison of the COOP charts and the SF-36. Pain 1999;83:201−9.
[48] Bullinger M. German translation and psychometric testing of the SF-36 health survey: preliminary results from the IQOLA project. International quality of life assessment. Soc Sci Med 1995;41:1359−66.
[49] Dunn KM, Jordan K, Croft PR. Does questionnaire structure influence response in postal surveys? J Clin Epidemiol 2003;56:10−6.
[50] Gatchel RJ, Polatin PB, Mayer TG, Robinson R, Dersh J. Use of the SF-36 health status survey with a chronically disabled back pain population: strengths and limitations. J Occup Rehabil 1998;8:237−46.
[51] Weigl M, Ewert T, Kleinschmidt J, Stucki G. Measuring the outcome of health resort programs. J Rheumatol 2006;33:764−70.
[52] Zwingmann C, Metzger D, Jäckel W. Short Form-36 Health Survey (SF-36): psychometrische analysen der deutschen Version bei Rehabilitanden mit chronischen Rückenschmerzen. Diagnostica 1998;44:209−19.
[53] Diaz-Arribas MJ, Fernandez-Serrano M, Royuela A, Kovacs FM, Gallego-Izquierdo T, Ramos-Sanchez M, et al. Minimal clinically important difference in quality of life for patients with low back pain. Spine 2017;42:1908−16.
[54] Luo X, George ML, Kakouras I, Edwards CL, Pietrobon R, Richardson W, et al. Reliability, validity, and responsiveness of the short form 12-item survey (SF-12) in patients with back pain. Spine 2003;28:1739−45.
[55] Turner JA, Fulton-Kehoe D, Franklin G, Wickizer TM, Wu R. Comparison of the Roland-Morris Disability Questionnaire and generic health status measures: a population-based study of workers' compensation back injury claimants. Spine 2003;28:1061−7.
[56] Garratt AM, Klaber Moffett J, Farrin AJ. Responsiveness of generic and specific measures of health outcome in low back pain. Spine 2001;26:71−7.
[57] Mueller B, Carreon LY, Glassman SD. Comparison of the EuroQOL-5D with the Oswestry Disability Index, back and leg pain scores in patients with degenerative lumbar spine pathology. Spine 2013;38:757−61.
[58] Soer R, Reneman MF, Speijer BL, Coppes MH, Vroomen PC. Clinimetric properties of the EuroQol-5D in patients with chronic low back pain. Spine J 2012;12:1035−9.
[59] Whynes DK, McCahon RA, Ravenscroft A, Hodgkinson V, Evley R, Hardman JG. Responsiveness of the EQ-5D health-related quality-of-life instrument in assessing low back pain. Value Health 2013;16:124−32.
[60] Berwick DM, Murphy JM, Goldman PA, Ware JE Jr, Barsky AJ, Weinstein MC. Performance of a five-item mental health screening test. Med Care 1991;29:169−76.
[61] Davies AR, Ware JE. Measuring health perceptions in the health insurance experiment. Santa Monica, CA: Rand Corporation; 1981.
[62] Stewart AL, Hays RD, Ware JE Jr. The MOS short-form general health survey. Reliability and validity in a patient population. Med Care 1988;26:724−35.
[63] Stewart AL, Ware JE. Measuring functioning and well-being: the medical outcomes study approach. Durham, USA: Duke University Press; 1992.
[64] Hunt SM, McKenna S, McEwen J, Williams J, Papp E. The Nottingham Health Profile: subjective health status and medical consultations. Soc Sci Med A 1981;15:221−9.
[65] Hunt SM, McKenna SP, McEwen J, Backett EM, Williams J, Papp E. A quantitative approach to perceived health status: a validation study. J Epidemiol Community Health 1980;34:281−6.

[66] PROMIS instrument development and validation scientific standards version 2.0. Available at: http://www.healthmeasures.net/images/PROMIS/PROMISStandards_Vers2.0_Final.pdf2013. Accessed November 14, 2017.
[67] Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005−2008. J Clin Epidemiol 2010;63:1179−94.
[68] Cella D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years. Med Care 2007;45:S3.
[69] DeWalt DA, Rothrock N, Yount S, Stone AA. Evaluation of item candidates: the PROMIS qualitative item review. Med Care 2007;45:S12.
[70] Boers M, Kirwan JR, Wells G, Beaton D, Gossec L, d'Agostino M-A, et al. Developing core outcome measurement sets for clinical trials: OMERACT filter 2.0. J Clin Epidemiol 2014;67:745−53.
[71] Idzerda L, Rader T, Tugwell P, Boers M. Can we decide which outcomes should be measured in every clinical trial? A scoping review of the existing conceptual frameworks and processes to develop core outcome sets. J Rheumatol 2014;41:986−93.
[72] Wilson IB, Cleary PD. Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes. JAMA 1995;273:59−65.
[73] Saragiotto BT, Maher CG, Yamato TP, Costa LO, Menezes Costa LC, Ostelo RW, et al. Motor control exercise for chronic non-specific low-back pain. Cochrane Database Syst Rev 2016;Cd012004.
[74] Teresi JA. Overview of quantitative measurement methods. Equivalence, invariance, and differential item functioning in health applications. Med Care 2006;44:S39−49.
[75] Janssen MF, Pickard AS, Golicki D, Gudex C, Niewada M, Scalone L, et al. Measurement properties of the EQ-5D-5L compared to the EQ-5D-3L across eight patient groups: a multi-country study. Qual Life Res 2013;22:1717−27.
[76] Bagraith KS, Strong J, Meredith PJ, McPhail SM. Rasch analysis supported the construct validity of self-report measures of activity and participation derived from patient ratings of the ICF low back pain core set. J Clin Epidemiol 2017;84:161−72.
[77] Bagraith KS, Strong J, Meredith PJ, McPhail SM. Self-reported disability according to the international classification of functioning, disability and health low back pain core set: test-retest agreement and reliability. Disabil Health J 2017;10:621−6.
[78] Cieza A, Stucki G, Weigl M, Disler P, Jackel W, van der Linden S, et al. ICF Core Sets for low back pain. J Rehabil Med (44 Suppl), 2004;69−74.
[79] Hill JC, Kang S, Benedetto E, Myers H, Blackburn S, Smith S, et al. Development and initial cohort validation of the Arthritis Research UK Musculoskeletal Health Questionnaire (MSK-HQ) for use across musculoskeletal care pathways. BMJ Open 2016;6:e012331.
[80] Lin CW, McAuley JH, Macedo L, Barnett DC, Smeets RJ, Verbunt JA. Relationship between physical activity and disability in low back pain: a systematic review and meta-analysis. Pain 2011;152:607−13.
[81] Maas ET, Juch JN, Ostelo RW, Groeneweg JG, Kallewaard JW, Koes BW, et al. Systematic review of patient history and physical examination to diagnose chronic low back pain originating from the facet joints. Eur J Pain 2017;21:403−14.
[82] Machado GC, Maher CG, Ferreira PH, Day RO, Pinheiro MB, Ferreira ML. Non-steroidal anti-inflammatory drugs for spinal pain: a systematic review and meta-analysis. Ann Rheum Dis 2017;76:1269−78.
[83] Parreira P, Heymans MW, van Tulder MW, Esmail R, Koes BW, Poquet N, et al. Back Schools for chronic non-specific low back pain. Cochrane Database Syst Rev 2017;8:Cd011674.
[84] Raastad J, Reiman M, Coeytaux R, Ledbetter L, Goode AP. The association between lumbar spine radiographic features and low back pain: a systematic review and meta-analysis. Semin Arthritis Rheum 2015;44:571−85.
[85] Wassenaar M, van Rijn RM, van Tulder MW, Verhagen AP, van der Windt DA, Koes BW, et al. Magnetic resonance imaging for diagnosing lumbar spinal pathology in adult patients with low back pain or sciatica: a diagnostic systematic review. Eur Spine J 2012;21:220−7.
[86] Chiarotto A, Clijsen R, Fernandez-de-Las-Penas C, Barbero M. Prevalence of myofascial trigger points in spinal disorders: a systematic review and meta-analysis. Arch Phys Med Rehabil 2016;97:316−37.